Abstract

In this document, we describe the statistical features of Scilab. We analyse the features available in Scilab's core (i.e. provided "out of the box") and Scilab Statistical Toolboxes. For Scilab's core statistical features, we analyse the different libraries used by Scilab and provide a complete overview of the functions. For the most important features, we present Scilab sessions with a sample use of the command. Several Scilab Toolboxes are analysed in this document, including Sci_R and Stixbox. We also analyse the missing features (not provided in the core and not in the toolboxes) with the tools which are available in other languages, including Matlab and R.

Scilab and Statistics

Michael Baudin

February 2009

Contents

1 Introduction **2**

1.1 A sample session 2

2 Scilab statistical features **6**

2.1 The sources 6

2.2 Overview of functions 6

2.2.1 General description functions 6

2.2.2 Special functions 6

2.2.3 Cumulated density functions 6

2.2.4 Random number generation 11

3 Statistical Toolboxes **12**

4 Missing features **13**

Bibliography **14**

Chapter 1

Introduction

As stated in [16], Scilab's core provide a complete set of features related to simulation and statistical computations. Indeed, Scilab provide uniform pseudo-random number generators, functions to compute the moments of a distribution and a complete set of distributions. In this document, we will make a complete overview on these features.

It must be noticed, though, that these features are not as complete as in other languages, like R for example. This is why several toolboxes have developped in order to extend the features of Scilab. In this document, we will present two major toolboxes, that is the Sci_R toolbox and the Stix toolbox.

In the last chapter, we will analyse the missing statistical features and will analyse how these features are available in other tools, such as Matlab, R, or Octave.

1.1 A sample session

A good introduction on the statistical features of Scilab is [5]. In the remaining of this introductin chapter, we will try to have a flavour of how to perform statistical computations with Scilab. We focus on the algorithms which are used inside Scilab, to show what exact algorithms perform the computations.

As a first example, we will generate a sequence of numbers from a normal law with mean 0 and standard deviation 1 (example inspired and simplified from [5]). The probability density function (pdf) and the cumulated probability density function of the normal law is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \tag{1.1}$$

$$P(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}. \tag{1.2}$$

The empirical cumulated density function [12] of a given set of data $\{x_i\}_{i=1,N}$ is given by

$$F_N(x) = \frac{\text{number of } x_1, x_2, \ldots, x_n \text{ that are } \leq x}{N}. \tag{1.3}$$

The numerical method used by Scilab to generate such numbers is the Polar method for normal deviates, as presented in [12].

```
N=200;
x = rand(1,N,"normal");
Xsorted =gsort(x,"g","i");
Ydata = (1:N)/N;
plot(Xsorted,Ydata);
e=gce();
e.children.polyline_style=2;
xtitle("Empirical Cumulated Density Function of Normal Law with 200 samples")
filename = "introduction_ecdfnormal.png";
xs2png(0,filename);
```

The empirical cumulated density function is presented in figure 1.1.

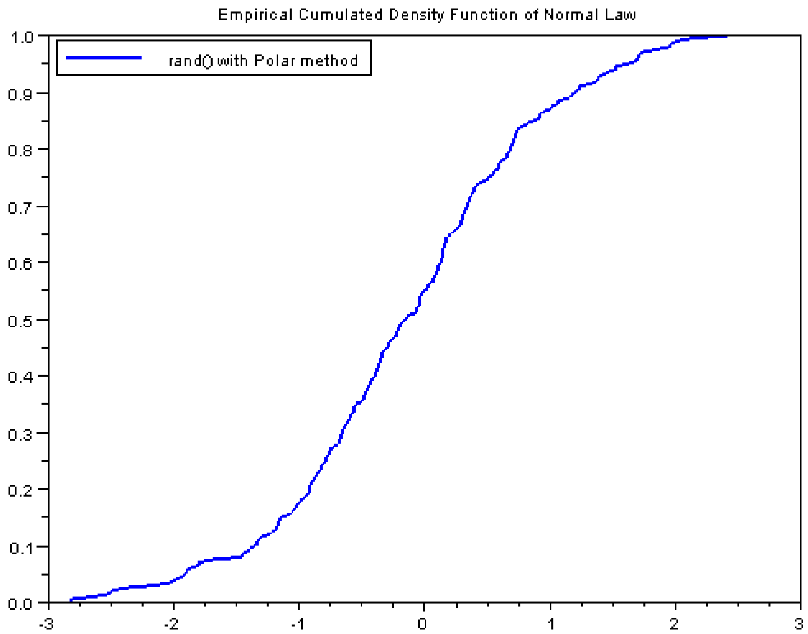

Figure 1.1: Empirical Cumulated Density Function of Normal Law with 200 samples

To compare the data which is produced by rand with the cumulated density function of the normal law, we use the *cdfnor* primitive. This primitive is based on [6] and uses rational functions that theoretically approximate the normal distribution function to at least 18 significant decimal digits. The same primitive can be used to compute the inverse of the cumulated density function. In that case, rational functions are used as starting values to Newton's Iterations which compute the inverse standard normal.

```
N=200;
x = rand(1,N,"normal");
Xsorted =gsort(x,"g","i");
Ydata = (1:N)/N;
x=linspace(-3,3,100);
P=cdfnor("PQ",x,zeros(x),ones(x));
plot(Xsorted,Ydata,x,P);
```

The comparison plot between the empirical cdf and the computed cdf is presented in figure 1.2.

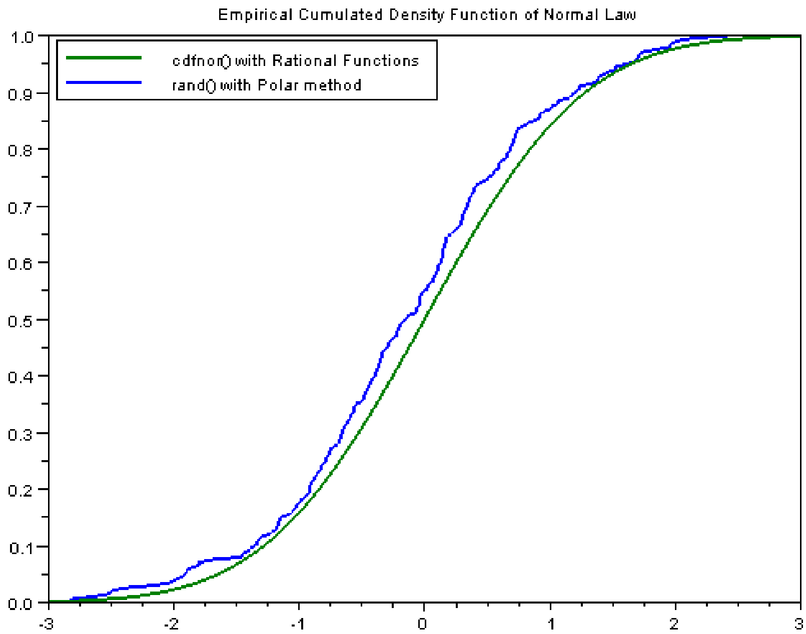

Figure 1.2: Cumulated Density Function of Normal Law : comparison of cdf from rational functions and empirical cdf from Polar method

The moments of a distribution can be computed with the *mean*, *variance* and *stdev* Scilab macros, which are implementations of the moments. For the variance and standard deviation, the scaling factor is $N-1$. In the following script, one computes these moment for an increasing number of samples, from 10^1 to 10^5.

```
nbpoints = 5;
means=zeros(nbpoints,1);
vars=zeros(nbpoints,1);
stdevs=zeros(nbpoints,1);
nlist = 1:nbpoints;
for i = nlist
  N=10^i;
  x = rand(1,N,"normal");
  means(i) = mean(x);
  vars(i) = variance(x);
  stdevs(i) = stdev(x);
end
plot(nlist,[means,vars,stdevs]);
```

The convergence plot of the moments is presented in figure 1.3.

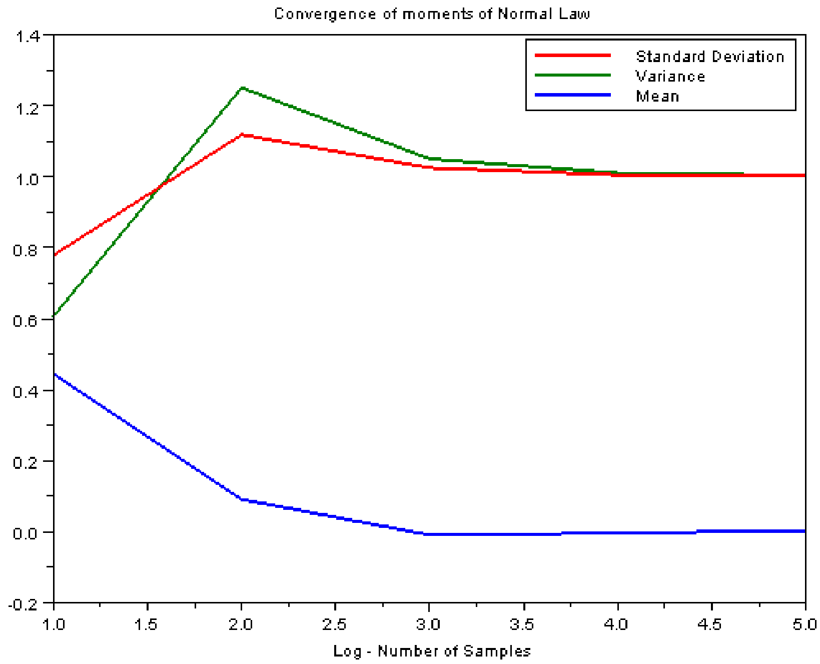

Figure 1.3: Convergence of the moments of the normal law

Chapter 2

Scilab statistical features

In this chapter, we describe the features which are provided in Scilab's core, that is, "out of the box". Indeed, Scilab provide features such as general statistical description of datas, many cumulated density functions and can generate uniform and non uniform random variates. These features are based on several open source libraries, that we are analysing in the first section. A complete overview of these features is provided in the second section, where we analyse the full list of functions and the numerical methods they use. For the most important functions, we provide a sample session where the function is used and some plots of the results.

2.1 The sources

In this section, we analyse the libraries which are available in Scilab and which provide its statistical features. The figure 2.1 is an overview of the libraries which are either Scilab macros or source code, provided in C, Fortran 77 or as Scilab macros.

2.2 Overview of functions

The figure 2.2 presents a complete list of Scilab statistical functions.

2.2.1 General description functions

The figure 2.3 presents the general description functions available in Scilab.

2.2.2 Special functions

The figure 2.4 presents the special functions available in Scilab.

The figure 2.5 presents a detailed analysis of the location and internal design of the special functions available in Scilab.

2.2.3 Cumulated density functions

The figure 2.6 presents the cumulated density functions available in Scilab.

Commands	calerf, erf, erfc, erfcx
Routines	CALERF
Directory	scilab/modules/elementary_functions/src/fortran
Language	Fortran
Download	`http://www.kurims.kyoto-u.ac.jp/~ooura/index.html`
Author	Takuya Ooura
Year	1996
References	[6]

Name	Labostat
Directory	scilab/modules/elementary_functions/src/fortran
Commands	General description functions (center, variance, etc...)
Language	Scilab scripts
Author	Carlos Klimann
Year	2000
References	[15], [14]

Name	DCDFLIB
Directory	scilab/modules/statistics/src/dcdflib
Download	`http://www.netlib.org/random/`
Commands	Cumulated Density Functions (cdfbet, cdfbin, etc...)
Language	Fortran
Author	Barry Brown, W. J. Cody, Alfred H. Morris Jr
Year	1994 for library, 1992 for code by Cody, 1991 for code by Morris
References	[1], [10], [6], [11] [9], [8]

Name	Randlib
Directory	scilab/modules/randlib/src/fortran
Download	`ftp://odin.mda.uth.tmc.edu/pub/source`
	(unavailable at the time of the writing of this report)
Commands	grand (for distributions like normal, gamma, chi, etc...)
Language	Fortran
Author	Barry Brown, James Lovato, Kathy Russell, John Venier
Year	1997
References	[2], [4], [7], [3]

Figure 2.1: Statistical libraries available in Scilab

Description of Data	
center	cmoment
correl	covar
ftest	ftuneq
geomean	harmean
iqr	labostat
mad	mean
meanf	median
moment	msd
mvvacov	nancumsum
nand2mean	nanmax
nanmean	nanmeanf
nanmedian	nanmin
nanstdev	nansum
nfreq	pca
perctl	princomp
quart	regress
sample	samplef
samwr	show_pca
st_deviation	stdevf
strange	tabul
thrownan	trimmean
variance	variancef
wcenter	

Special Functions	
beta	calerf
erf	erfc
erfcx	erfinv
gamma	gammaln

Random Number Generation	
grand	prbs_a
rand	sprand
randpencil	

Cumulated Density Functions	
cdfbet	cdfbin
cdfchi	cdfchn
cdff	cdffnc
cdfgam	cdfnbn
cdfnor	cdfpoi
cdft	

Figure 2.2: Complete list of statistical features in Scilab

Name	Feature
center	center
wcenter	center and weight
cmoment	central moments of all orders
correl	correlation of two variables
covar	covariance of two variables
ftest	Fischer ratio
ftuneq	Fischer ratio for samples of unequal size.
geomean	geometric mean
harmean	harmonic mean
iqr	interquartile range
mad	mean absolute deviation
mean	mean (row mean, column mean) of vector/matrix entries
meanf	weighted mean of a vector or a matrix
median	median (row median, column median,...) of vector/matrix/array entries
moment	non central moments of all orders
msd	mean squared deviation
mvvacov	computes variance-covariance matrix
nancumsum	Thos function returns the cumulative sum of the values of a matrix
nand2mean	difference of the means of two independent samples
nanmax	max (ignoring Nan's)
nanmean	mean (ignoring Nan's)
nanmeanf	mean (ignoring Nan's) with a given frequency.
nanmedian	median of the values of a numerical vector or matrix
nanmin	min (ignoring Nan's)
nanstdev	standard deviation (ignoring the NANs).
nansum	Sum of values ignoring NAN's
nfreq	frequence of the values in a vector or matrix
pca	Computes principal components analysis with standardized variables
perctl	computation of percentils
princomp	Principal components analysis
quart	computation of quartiles
regress	regression coefficients of two variables
sample	Sampling with replacement
samplef	sample with replacement from a population and frequences of his values.
samwr	Sampling without replacement
show_pca	Visualization of principal components analysis results
st_deviation	standard deviation (row or column-wise) of vector/matrix entries
stdevf	standard deviation
strange	range
tabul	frequency of values of a matrix or vector
thrownan	eliminates nan values
trimmean	trimmed mean of a vector or a matrix
variance	variance of the values of a vector or matrix
variancef	standard deviation of the values of a vector or matrix

Figure 2.3: Description of Data functions

Name	Feature
beta	beta function
calerf	computes error functions
erf	error function
erfc	complementary error function
erfcx	scaled complementary error function
erfinv	inverse of the error function
gamma	gamma function
gammaln	logarithm of gamma function

Figure 2.4: Special functions

Name	Location / Internals
beta	modules/special_functions/sci_gateway/c/sci_beta.c switch to dgammacody by W. J. Cody and L. Stoltz and to betaln from DCDFLIB
calerf	modules/elementary_functions/src/fortran by Takuya OOURA
erf	modules/elementary_functions/macros/erf.sci call to calerf
erfc	modules/elementary_functions/macros/erfc.sci call to calerf
erfcx	modules/elementary_functions/macros/erfcx.sci call to calerf
erfinv	modules/special_functions/macros/erfinv.sci rational aproximation of erfinv + 2 Newton's steps
gamma	modules/special_functions/sci_gateway/fortran/sci_f_gamma.f based on dgammacody by W. J. Cody and L. Stoltz
gammaln	modules/elementary_functions/src/fortran/dlgama.f by W. J. Cody and L. Stoltz

Figure 2.5: Detailed analysis of special functions

Name	Feature
cdfbet	Beta distribution
cdfbin	Binomial distribution
cdfchi	chi-square distribution
cdfchn	non-central chi-square distribution
cdff	F distribution
cdffnc	non-central F distribution
cdfgam	gamma distribution
cdfnbn	negative binomial distribution
cdfnor	normal distribution
cdfpoi	poisson distribution
cdft	Student's T distribution

Figure 2.6: Cumulated density functions

2.2.4 Random number generation

The figure 2.7 presents the random number generators available in Scilab.

Name	Feature
grand	Random number generators
prbs_a	pseudo random binary sequences generation
rand	random number generator
sprand	sparse random matrix
randpencil	random pencil

Figure 2.7: Random number commands

The figure 2.7 presents a detailed analysis of the location and design of the random number generators available in Scilab.

Name	Location / Internals
grand	modules/randlib/sci_gateway/c/sci_grand.c
	based on several random number generators
prbs_a	modules/cacsd/macros/prbs_a.sci
	based on rand
rand	modules/elementary_functions/src/fortran/urand.f
	by Michael A. Malcolm And Cleve B. Moler
sprand	(todo)
randpencil	(todo)

Figure 2.8: Detailed analysis of random number commands

Chapter 3

Statistical Toolboxes

http://www.scilab.org/contrib/index_contrib.php?page=download&category=DATA%20ANALYSIS%
20AND%20STATISTICS

GLMBOX :generalized statistical linear models analysis. (Dec 2003). http://www.scilab.org/contrib/index_contrib.php?page=displayContribution&fileID=183

grocer 1.2 : Comprehensive econometric toolbox http://www.scilab.org/contrib/index_contrib.php?page=displayContribution&fileID=248

Hurst : Exponent estimators v2.0 http://www.scilab.org/contrib/index_contrib.php?page=displayContribution&fileID=988

multilinear regression http://www.scilab.org/contrib/index_contrib.php?page=displayContri
339

Sci_R for scilab 5.x http://www.scilab.org/contrib/index_contrib.php?page=displayContribu
1138

stixbox 1.2.5 http://www.scilab.org/contrib/index_contrib.php?page=displayContribution&
184 statistics toolbox designed for the french examination "agregation de mathematiques"

Chapter 4

Missing features

- Empirical Cumulated Density Function
- Robust implementation of variance, standard deviation. See in ”Art of Computer Programming” [12], chapter 4.2.2, ”Accuracy of Floating Point Arithmetic”, section A or in ”Numerical Recipes” [13], chapter 14.1, ”Moments of a Distribution: Mean, Variance, Skewness, and so Forth”.

Bibliography

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[2] J. H. Ahrens and U. Dieter. Computer methods for sampling from the exponential and normal distributions. *Commun. ACM*, 15(10):873–882, 1972.

[3] J. H. Ahrens and U. Dieter. Extensions of forsythe's method for random sampling from the normal distribution. *Mathematics of Computation*, 27(124):927–937, 1973.

[4] J. H. Ahrens and U. Dieter. Generating gamma variates by a modified rejection technique. *Commun. ACM*, 25(1):47–54, 1982.

[5] J.-P. Chancelier, F. Delebecque, C. Gomez, M. Goursat, R. Nikoukhah, and S. Steer. *Introduction à Scilab.* Springer, 2007.

[6] W. J. Cody. Algorithm 715: Specfun–a portable fortran package of special function routines and test drivers. *ACM Trans. Math. Softw.*, 19(1):22–30, 1993.

[7] Luc Devroye. Non-uniform random variate generation, 1986. http://cg.scs.carleton.ca/~luc/rnbookindex.html.

[8] Armido R DiDonato and Alfred H Morris, Jr. Computation of the incomplete gamma function ratios and their inverse. *ACM Trans. Math. Softw.*, 12(4):377–393, 1986.

[9] Armido R. Didonato and Alfred H. Morris, Jr. Algorithm 708: Significant digit computation of the incomplete beta function ratios. *ACM Trans. Math. Softw.*, 18(3):360–373, 1992.

[10] John F. Hart, E. W. Cheney, Charles L. Lawson, Hans J. Maehly, Charles K. Mesztenyi, John R. Rice, Thacher, and Christoph Witzgall. *Computer Approximations.* SIAM Series on Applied Mathematics. John wiley & Sons, 1968.

[11] William J. Kennedy and James E. Gentle. *Statistical Computing.* Marcel Dekker, Inc., New York, 1980.

[12] D. E. Knuth. *The Art of Computer Programming, Volume 2, Seminumerical Algorithms.* Third Edition, Addison Wesley, Reading, MA, 1998.

[13] W. H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C, Second Edition.* 1992.

[14] Gilbert Saporta. *Probabilités, analyses des données et statistiques.* Editions Technip, June.

[15] T.H. Wonacott and R.J. Wonacott. *Introductory statistics.* John Wiley and Sons, New York, 1990.

[16] Bernard Ycart. Démarrer en scilab suivi de statistiques en scilab. http://ljk.imag.fr/membres/Bernard.Ycart/polys/demarre_scilab/node21.html, 2001.